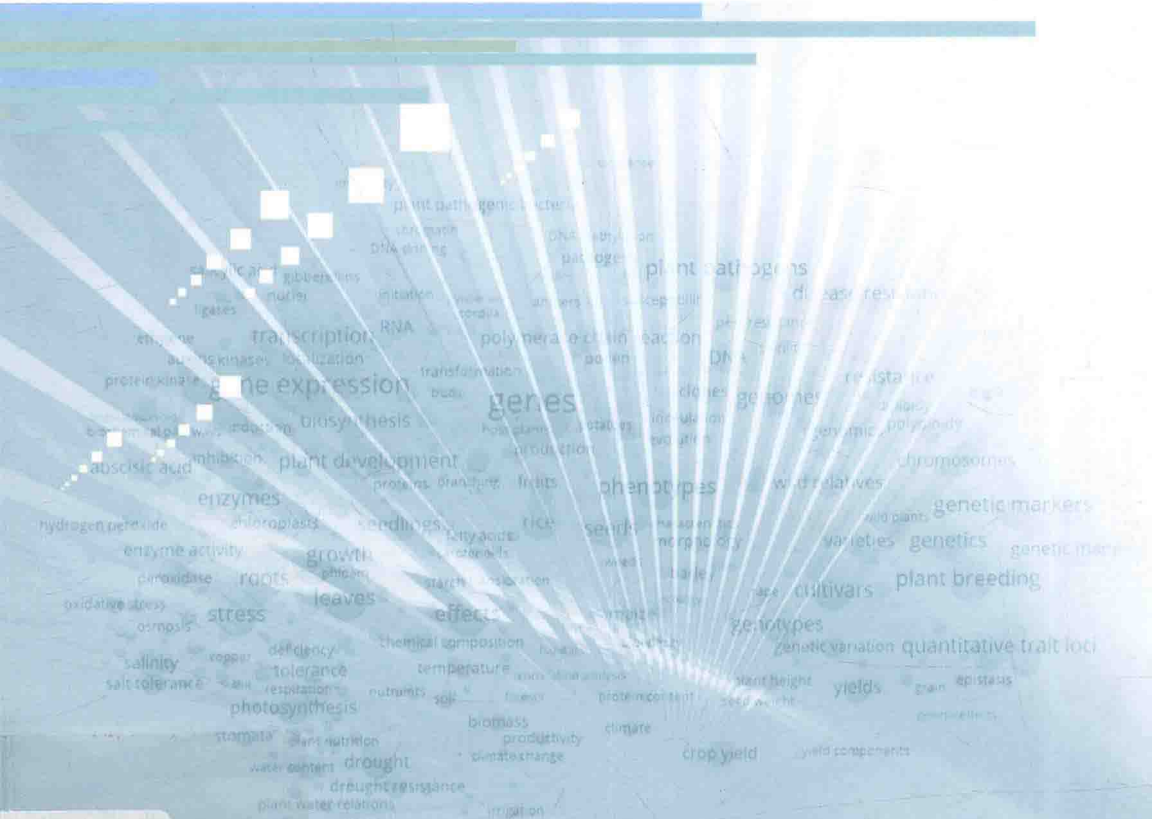
· 刘敏娟
· 张学福 著

基于知识图谱的学科主题识别方法研究



中国农业科学技术出版社

· 刘敏娟
· 张学福 著

基于知识图谱的学科主题识别方法研究

中国农业科学技术出版社

图书在版编目（CIP）数据

基于知识图谱的学科主题识别方法研究 / 刘敏娟，张学福著 . —北京：中国农业科学技术出版社，2017. 3

ISBN 978 - 7 - 5116 - 2835 - 0

Ⅰ. ①基… Ⅱ. ①刘…②张… Ⅲ. ①科技情报 - 情报检索 - 研究 Ⅳ. ①G252. 7

中国版本图书馆 CIP 数据核字（2016）第 275571 号

责任编辑 穆玉红

责任校对 马广洋

出 版 者 中国农业科学技术出版社

北京市中关村南大街 12 号 邮编：100081

电 话 (010)82106626(编辑室) (010)82109702(发行部)

(010)82109709(读者服务部)

传 真 (010)82106626

网 址 http://www.castp.cn

经 销 者 各地新华书店

印 刷 者 北京富泰印刷有限责任公司

开 本 710mm ×1 000mm 1/16

印 张 8.5

字 数 120 千字

版 次 2017 年 3 月第 1 版 2017 年 3 月第 1 次印刷

定 价 38.00 元

前　言

基于知识图谱的学科主题识别方法研究是结合多种计量学方法与科学知识图谱技术，深入研究分析学科知识体系的结构关系，辨识和探测学科领域的研究热点主题及其变化趋势的方法研究，以期帮助科研人员更好地从大规模科技文献中迅速掌握学科结构与热点主题，成为新环境下科技决策者有效开展科技管理工作的新手段和新途径。

本研究主要包括以下 4 个方面：①对学科主题识别理论及方法加以总结、梳理和提炼，改进了现有的主题识别分析流程，为后期研究奠定了理论基础。②研究了学科领域分析数据集构建的方法。创新性地提出了一种基于期刊关系的构建学科领域分析数据集的方法，包括分析模型与分析流程的构建以及期刊关系分析中涉及的关键技术研究。③探讨了学科主题识别的方法。对关键技术环节的处理方法进行了研究改进，创新性地提出了一种基于词频、词量、累积词频占比三者关系，依据不同词集实际分布规律和特点，动态确定高、中、低频词的阈值，划定高、中频词作为共词分析的词集范围的方法。同时还创新性地提出了一种基于核心词、突变词、新生词对领域中核心主题、一般主题、新兴主题的变化趋势进行追踪的方法，综合运用了词频分析、共词分析、聚类分析和知识图谱多种手段，最终达到学科领域主题变化趋势识别的目的。④以作物学科为例验证了学科主题识别方法的可行性。以作物学科为例，重点对前述方法进行了实证分析，最终结果得到了作物学领域专家的认可，并与常用方法进行了对比分析，验证了本研究方法的可行性及效果。

本研究通过理论分析、方法研究与实证分析，验证了本研究中提出的方法的科学性与有效性，拓展了学科主题识别的理论与方法，为引入科学知识图谱技术进行学科领域分析数据集构建以及学科主题识别的深入研究提供了有意义的参考和借鉴。

作　者

2016. 10

目　录

图目录

表目录

第一章 绪 论

第一节 研究背景

科技文献作为科学技术的重要软载体形式，其中蕴含着大量的科学技术知识，也是科研技术成果发表、溯源、引证和查新的重要来源，是科研人员获取相关信息与知识的最重要途径之一。但如今，现代科学技术的突飞猛进，并伴随国际互联网络发展而在世界上的迅速传播，导致全球知识呈爆炸式增长，由此也带来了知识与信息选择的困难。科研人员面对着呈爆炸式增长的海量文献数据与信息知识的复杂多变，借助传统的经典文献检索方法来查询重要文献获取专业知识，已难以保证查找到关键性文献和前沿知识，无法从海量文献集合中快速获取知识、把握科研动态。因此，对科研人员来说，如何快速、准确地掌握学科领域的发展动态，了解研究热点主题，找到科技创新的突破口，已经成为他们的迫切需求。同样，传统的科技管理理论和方法在新时期也遇到诸多问题。科技管理人员和科技政策的制定者需要站在全球高度，从宏观、整体、顶层角度全面了解领域科技现状与最新态势，紧跟科技发展的步伐，对各个领域未来的科技发展趋势做出准确预判。但是，瞬息万变的知识更新节奏，海量的科技信息数据，使科技决策人员难以及时甄别并获得有价值的信息和知识。因此，迫切需要新的科技评价手段帮助科技管理人员突破现有科技评价模式的瓶颈，提高

科技决策的效率。

学科情报研究工作是针对特定主题，收集、积累相关文献、数据等信息，并加以整理、分析和研究，最终根据用户的需要提出分析研究结果或报告的全过程（徐敏等，2005）。情报分析人员可以通过对学科主题的结构及变化情况的识别，在纷繁复杂的学科信息中揭示学科知识发展变化及其相互作用的特征与规律，追踪学科的变化发展，确定该领域新出现的热点问题，探测领域的未来发展方向等，这无论是在战略或战术层面都可以提供有价值的情报和决策支持。进而，科研人员与科技管理人员可以利用学科领域的热点主题、研究前沿和焦点演变来寻找新的科研突破口，为宏观战略决策的制定提供重要参考。

当前，科学知识图谱（Mapping Knowledge Domains）作为一种追踪学科发展动态、探测学科知识结构、识别领域学科热点主题的新手段正在蓬勃兴起。它把现代科学技术知识的复杂领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，利用科学知识图谱理论与方法，可以针对某一产业或者学科专业领域的科学技术发展态势及其相关知识结构进行深入研究，进而发现科技活动中潜在的一般和特殊规律，使研究人员对于如何选择感兴趣的新领域也不再困难，也使得科技管理人员可以有效监测学科发展动态，从而推动学科整体的发展。

基于科学知识图谱的学科主题识别方法是由一套完整的分析流程和方法体系构成的对学科领域主题进行辨识的综合分析方法，包括文献计量方法、内容分析法、多元统计分析方法、社会网络分析法。如今，科学知识图谱与其它方法相结合已经广泛应用到学科领域主题识别的研究当中，然而这些方法自身的研究中尚存在一定的缺陷，而且方法组合策略以及如何组合运用这些方法进行主题识别的过程仍需要进一步优化和规范，其中涉及的一些关键环节也有进一步改进的空间。

基于以上研究背景，本研究将开展基于知识图谱的学科主题识别

方法研究，以期更好地帮助科研人员更好地从大规模科技文献中迅速掌握学科结构与热点主题，成为新环境下科技决策者有效开展科技管理工作的新手段和新途径。

第二节 国内外研究现状

学科领域主题识别，是对某一学科或某一研究领域的主题结构、研究热点、新兴主题以及主题变化趋势进行发现和识别的过程。目前主要有两类方式：一是定性研究，是以领域专家的经验知识为依托，充分发挥出专家在该领域的辨识力，定性地分析出学科或研究领域的状况及趋势、领域中出现的新兴研究主题和研究热点，但这种方法容易受所选专家自身知识及专家主观性的限制影响；二是定量研究，是通过计算机手段对科学文献数据进行定量的统计和分析来揭示和预测研究领域的主题结构和发展趋势，以计量学方法为主，但近年来随着科技的不断发展，社会网络分析、信息可视化等方法和技术也逐渐应用到学科领域分析中。

以下将分别从两个角度对相关文献进行综述归纳：一是对不同方法在学科领域识别中应用的国内外现状进行归纳；二是对学科领域识别中涉及的关键环节所涉及方法的国内外研究现状进行综述。由于本研究重点是定量化方法对学科主题的识别研究，因此综述内容将侧重定量化方法的相关研究。

一、不同方法应用于学科主题识别的相关研究

基于计量方法进行学科领域识别的主要方法包括共引分析法、共词分析法、词频分析法，同时当前研究人员也将这些方法与科学知识图谱技术相结合进行了广泛的应用。以下将分别对共引分析法、共词

分析法、词频分析法以及这些方法与科学知识图谱相结合在学科主题识别中的应用现状进行综述。

（一）共引分析法的应用

1973 年，“共引”（Co - citation）的概念由美国情报学家 Henry Small 提出，他指出当两篇不同文献共同被其它文献引用时，就说明这两篇文献存在共引关系，而且同被引的文献在主题上是具有相似性的，同时又提出了“共引强度”（Co - citation Strength）的概念，将其定义为两篇文献同时被其它文献引用的次数，这一概念主要用来测量文献在内容主题上的相关程度，即共引强度越高，文献之间的相关度越高（Small H，1973）。共引的概念提出以后，国外采用共引分析进行了深入的研究并广泛应用。1973 年，Small 采用共引分析方法测度了粒子物理学领域高被引文献之间的相关关系，对比了引文网络的差异，研究了文献结构的变迁（Small H，1973）。随后，Small 又对共引分析进行了深入研究。2006 年，Small 对 ESI 高被引论文进行共引聚类，追踪不同时间片共引网络中共引论文簇的变化，从而绘制论文簇的发展路线图，研究了全学科领域的科学结构（Small H，2006）。2009 年，Small 又基于共引分析方法研究了有机薄膜传感器领域从新突现、发展应用到消亡过程中引文网络的结构属性的变化（Small H，2009）。2010 年，日本 NISTEP 研究所对 ESI 的前 1% 的高被引论文进行聚类，绘制了 22 个学科领域的科学结构地图，并探测其研究前沿（Ayaka S，2010）。国内，近几年主要的共引分析实证研究有：刘泽渊，王贤文对 1975—2007 年收录生态经济学论文的引文数据进行文献共被引分析，识别出生态经济学的主要知识群，研究了生态经济学 3 个不同的阶段：学科基础形成、学科领域拓展、学科纵深发展（刘则渊等，2008）。2009 年，王小梅等通过对 2002—2007 年 ESI 的高被引论文进行同被引聚类，分析了 121 个研究领域的学科交叉性，提供了一种对科学结构及

其演化进行研究和揭示的方法（王小梅等，2009）。2010 年，黄维、陈勇运用文献共被引分析，对2000—2008 年与教育经济学领域相关的6 个学术期刊发表的2 006篇论文的26 753条引文进行多维尺度分析，揭示出我国教育经济学领域的主题结构及研究热点（黄维等，2010）。2013 年，连少华、王宇综合运用频次统计论文同被引等文献计量分析方法，同时将因子分析和社会网络分析运用其中，研究了过去15 年我国图书馆学研究领域的研究主题分布状况，并揭示该领域的研究热点，预测了我国图书馆学未来的发展趋势（连少华等，2013）。王俏等结合共被引关理论，分析医学信息学相关学科分布情况，归纳其分布特征，旨在理清学科发展历程，为相关研究者提供借鉴（王俏等，2014）。发现共引分析是探测科学领域发展的有效工具，而且能测度特定领域之间的相关度，使用共引分析方法揭示科学结构及其演化。

（二）共词分析法的应用

共词分析法已经被科研人员进行了深入的研究，因其具有操作简单且结果易于解读的优点，使其在学科主题的识别中产生了较为广泛的应用。国外对共词分析应用的领域十分广泛，在专利文献、化学工程、软件工程、神经网络研究、生物安全、生物技术、环境工程、帕金森病、社会科学等领域都进行了分析，识别了相关领域的主题（Courtial J P 等，1993；Peters H P 等，1993；Van Raan，A. F. J.，1993；Noyons，E. C. M.；1998；Ho，Y. S，2007；Kostoff R. N，2009）。2001 年，Ding 基于共词分析法分析了之前30 年间信息检索领域研究主题发展变化的情况（Ding Y，2001）。2008 年，Lee 对科技文本进行了共词分析，从而发现了最新研究主题的变化趋势（Lee W. H.，2008）。国外研究不仅应用领域广泛，而且多学科领域的研究人员均参与到相关研究中去，从而促进了共词分析方法在学科领域主题识别研究中的迅速推广应用。不同的是，共词分析法的应用在国内仍处于推广阶段，

图情领域人员为主要研究力量，因此应用的研究领域也多集中于图情领域。2010 年，杨颖运用共词分析研究揭示了医学领域的学科结构（杨颖，2010）。2011 年，李颖利用共词分析对国内竞争情报领域的研究主题演变态势进行了分析（李颖，2011）。2009 年，邱均平对利用共词分析对我国图情学领域进行了分析，梳理了我国图情专业研究的重点主题及方向，并对未来的研究趋势进行了前瞻性的思考（邱均平等，2009）。2008 年，李长岭采用共词分析方法分析了 2002—2006 年图情专业的研究热点，主要利用该专业硕士论文中的高频关键词进行了共词聚类，在聚类比较的结果上比较研究热点的不同（李长岭等，2008）。2013 年，皇甫青红利用共词分析和社会网络分析方法，对国际数字图书馆领域的文献关键词进行因子分析和聚类分析，探讨该领域的数字图书馆虚拟技术研究、资源组织研究、资源建设研究、电子资源及版权研究和信息服务研究 5 大研究主题并找到各个研究主题内的作者团体（皇甫青红等，2013）。2014 年，沈同平、杨松涛等利用 SATI 3.2 软件统计 CNKI 全文数据库中知识链领域所涉及的关键词的词频，确定我国知识链研究领域使用频率最高的 32 个关键词，然后构造共词矩阵、相关矩阵、相异矩阵，并对不同的矩阵进行共词分析。最后，对分析结果进行讨论，归纳国内知识链理论研究热点（沈同平等，2014）。

（三）词频分析法的应用

词频分析法是文献计量方法的一种，主要通过计算某一研究领域文献中能够反映其核心内容的关键词或主题词的词频，来发现该领域研究热点主题及发展趋势的方法，即认为在领域文献中出现频次高的关键词或主题词可以代表该领域的研究热点主题。

目前词频分析法在学科领域主题识别中已经得到广泛应用。例如，1997 年，Robert 对 79 个纳米科技关键词进行词频分析，对世界纳米科

技研究状况进行了计量分析，揭示了全球范围内纳米科技论文的产出和分布（Robert D.，1997）。2003 年，梁立明采用词频分析法研究了我国纳米科技发展的情况（梁立明，2003）。2003 年，邱均平等运用词频分析法研究了国内外情报学研究热点主题和发展趋势（邱均平等，2003）。2006 年，马费成等人同样利用词频分析法揭示了国内外知识管理领域的研究热点及国内外的差异（马费成等，2006）。

（四）科学知识图谱在学科主题识别的应用

科学知识图谱是现实科学知识的发展与结构关系的一种图形，是科学计量学具有前景的研究方向（陈悦，2005）。科学知识图谱具有“图”和“谱”的双重性质和特征，既是可视化的知识图形，又是序列化的知识谱系，显示了知识元或知识群之间的网络、结构、互动、交叉、演化或衍生等诸多复杂关系（刘则渊，2008）。常见的科学知识图谱是聚类图谱、战略坐标图谱、多维尺度分析图谱、社会网络图谱。国内外很多学者运用共引分析、共词分析、词频分析方法结合科学知识图谱技术对不同领域进行学科主题的识别。

随着电子信息技术的迅速发展，人们获取大量信息后不再停留在数据层面，而是采用各种先进的可视化技术将其绘制成科学知识图谱。尤其是 20 世纪 90 年代以来科学知识图谱得到了迅猛发展。

早在 1964 年，加菲尔德就曾手工绘制了揭示科学引文演进的科学引文编年图谱，旨在探测科学在发展过程中重大事件发生的背景、发展概貌、突破性成就等。1973 年，Small 通过科学知识图谱探测了科学范式的演进（SmalI，1973）。1998 年，White 绘制了信息科学领域的科学知识图谱，界定了 1972—1995 年的信息科学的 12 个分支学科（White，1998）。1998 年，E. C. M. Noyons 等绘制了科学知识图谱识别出 1992—1997 年科学计量学、信息计量学及文献计量学的 5 个主要分支领域（Noyons 等，1998）。2001 年，Garfield 等以 SCI 等数据库为数